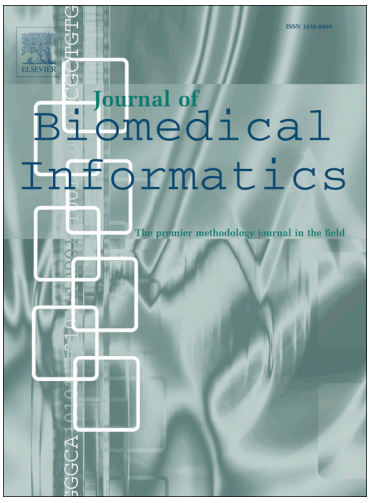


An Empirical Study on Prediction of Population Health through Social Media

Hung Nguyen, Thin Nguyen, Duc Thanh Nguyen

PII: S1532-0464(19)30196-0
DOI: https://doi.org/10.1016/j.jbi.2019.103277
Reference: YJBIN 103277

To appear in: *Journal of Biomedical Informatics*

Received Date: 4 October 2018
Revised Date: 16 August 2019
Accepted Date: 28 August 2019

Please cite this article as: Nguyen, H., Nguyen, T., Nguyen, D.T., An Empirical Study on Prediction of Population Health through Social Media, *Journal of Biomedical Informatics* (2019), doi: https://doi.org/10.1016/j.jbi.2019.103277

An Empirical Study on Prediction of Population Health through Social Media

Hung Nguyen[a,c], Thin Nguyen[a], Duc Thanh Nguyen[b,*]

[a]*Deakin University,*
Applied Artificial Intelligence Institute, Waurn Ponds VIC 3216, Australia
[b]*Deakin University,*
School of Information Technology, Burwood VIC 3125, Australia
[c]*Nha Trang University,*
Faculty of Information Technology, Khanh Hoa, Vietnam

Abstract

Public health measurement is important for government administration as it provides indicators and implications to public healthcare strategies. The measurement of health status has been traditionally conducted via surveys in the forms of pre-designed questionnaires to collect responses from targeted participants. Apart from benefits, traditional approach is costly, time-consuming, and not scalable. These limitations make a major obstacle to policy makers to develop up-to-date healthcare programs. This paper studies the use of health-related information conveyed in user-generated content from social media for prediction of health outcomes at population level. Specifically, we investigate linguistic features for analysing textual data. We propose the use of visual features learnt from deep neural networks for understanding visual data. We introduce collective social capital information from location-based social media data. We conducted extensive experiments on large-scale datasets collected from two online social networks: Foursquare and Flickr, against the task of prediction of the U.S. county health indices. Experimental results showed that visual and collective social capital data achieved comparable prediction performance and outperformed textual information. These promising results also suggest the potential of social media for health analysis at population scales.

*Corresponding Author
Email addresses: hung@deakin.edu.au (Hung Nguyen), thin.nguyen@deakin.edu.au (Thin Nguyen), duc.nguyen@deakin.edu.au (Duc Thanh Nguyen)

Keywords: Public health, healthcare, social media

1. Introduction

Public health measurement is crucial for policy makers to estimate health outcomes, identify health-related risks and develop public healthcare programs for communities. The measures should reflect the communities' dynamic state of physical, mental and social well-being [1, 2]. Ideally, a useful public health measure should have the following characteristics: (1) ability to detect changes in health status of a population overtime, (2) validity of changes, (3) sensitivity to major health policy changes and (4) reliability and stability across various settings [3].

Traditional approach for population health measurement has been conducted via surveys. Surveyed data is commonly accumulated from the public through either telephone interviews or postal surveys in the forms of pre-designed questionnaires. The advantages of this approach are twofold: (1) the capability of obtaining targeted results as the questionnaires are pre-designed by experts, and (2) samples can be targeted and selected thoroughly. However, organising such surveys is usually expensive and time-consuming. Surveys often take years to make results publicly available. For instance, the Behavioral Risk Factor Surveillance System (BRFSS) 2017 reports[1] were created from the data collected in or before 2015. Furthermore, these surveys might not properly reflect the health status of the whole population due to the limited number of participants. For instance, BRFSS - the largest health survey conducted in the world - has been made from only about 400,000 interviews each year.

Social networks have been considered the backbone of social and economic life [4]. The advent of online social networking platforms, also known as "social media", has opened an abundant source of information for many healthcare applications [5]. Billions of people are daily connected via social networks such as Facebook, Twitter, Instagram, Flickr, or Foursquare. Recent developments in the Internet and mobile technologies have enriched social media to a higher level of socialisation by allowing

[1] https://www.cdc.gov/brfss/

people, easily and seamlessly, share their status, photos, activities, events and opinions with their connected friends or the public. As such, to some extent, the data from social media reflects health status, both mental and physical, of a population via mood, thinking, activities and communications of people in that population [6, 7].

Social media data offers a variety of data types, e.g., text, images, etc. Recently, *geocoded social media* records have been exploited as an observation proxy of real-world phenomena for health-related services [8]. Geocoded social media data can be used in many applications ranging from disaster management, routing services, and location-oriented health services. For instance, local events and venues of interest can be recommended based on the spatio-temporal information from geocoded data [9]. In [8], socio-spatial information from social media data was used to detect social processes, e.g., identifying regions by latent attitudes from Twitter posts.

Literature has also shown considerable associations between many health indices and *social capital*. For instance, Phung et al. [10] and Nguyen et al. [11] effectively established the link between online social capital calculated from individual social participation (number of groups, posts or comments) and social support (number of friends and followers, received responses) and their mood.

Although population health analysis using social media data has been studied extensively, existing works address various health-related issues on small and diversified datasets while there is lack of a comprehensive and conclusive overview on the field. This leads to difficulty in making suggestions for policy makers and health informatics researchers. To address these issues, we aim to provide a comprehensive investigation of population health analysis using social media data and perform an empirical study on existing approaches. Specifically, the contributions of our work include:

- We investigate various features extracted from social media data types. In particular, we investigate common linguistic features for analysing textual data. We propose to extract visual features using deep neural networks. We introduce collective social capital features from location-based social media data.

- We thoroughly evaluate social media data types in the task of prediction of population health outcomes. We conduct extensive experiments on large-scale

datasets collected from two location-centric social networks: Foursquare and Flickr. The health outcomes are estimated based on health-related questions in the Behavioral Risk Factor Surveillance System (BRFSS), which has been annually conducted by the U.S. Centers for Disease Control and Prevention and is currently the world largest health survey. To our understanding, this study is the first to benchmark different social media data types (especially multimedia data) for population health prediction on realistic and large-scale datasets.

2. Related Work

Social media has been the most popular means which allows people from everywhere to communicate and share their opinions, activities, and statuses. It hence has become a useful sensor for many healthcare applications. Understanding population health status from social media forms a new field namely "digital epidemiology". Online content, when harnessed appropriately, can also provide useful information about diseases and health dynamics in populations [12]. Literature has shown many Internet- and social media-based population health analysis studies. These works demonstrate the capability of social media in delivering an efficient yet reliable approach for monitoring health issues of populations.

2.1. Textual data

Textual posts have been considered as a primary data source for many health analysis studies. For instance, Google Flu Trends [13] was developed for analysing textual queries from web users on Google to track influenza-like illness in populations. Based on the frequency of requests on physician visits for influenza-like symptoms, the level of weekly influenza activities in each region in the U.S. could be estimated. Aramaki et al. [14] addressed a similar problem of detecting influenza epidemics but on Twitter posts. In this work, a support vector machine (SVM) [15] was employed to classify tweets that mentioned influenza patients. Experimental results showed that nearly 90% of the classification correlated to the gold standard. Furthermore, the work also demonstrated that Twitter posts could reflect the real world and that natural language processing (NLP) techniques could be applied to understanding the tweets. De Choudhury et

al. [16] explored the potential of social media as a measurement tool in understanding depression in populations. Specifically, a probabilistic model was developed to analyse Twitter posts made by individuals diagnosed with clinical depression to determine if the posts could indicate depression. The model took advantage of social activity, emotion and language manifested as signals reflecting individual mental health status. Extended from this model, "social media depression index" was proposed in [16] to characterise levels of depression in populations. Paul and Dredze [17] applied the Ailment Topic Aspect Model (ATAM) proposed in [18] on over 1.5 million health-related tweets and discovered the ailment-related behaviours. In this work, prior knowledge was also incorporated into several tasks including localising illnesses by geographic region and measuring behavioural risk factors. In [19, 7], textual post from Twitter [19] and Instagram [7] were used to understand the dietary choices and habits of populations. In [20], a sparse model was proposed to handle short and noisy text from social media data to detect wellness events including diet, exercise, and health.

2.2. Visual data

Visual data has recently been exploited as an important source of information that potentially captures health issues. For instance, Manikonda and De Choudhury [21] employed histograms of frequency of colours in imagery media data for understanding health-related issues and found that a large number of images (48-75%) from disorders contain highly saturated colours. In [22], hue, saturation, and value (HSV) of colours were used to analyse Instagram photos and considered as early screen markers of depression.

Although the capability of colour histograms and HSV in health analysis have been proved, those features are hand-crafted and thus require domain knowledge and manual design. In addition, those features are not able to capture high-level information (e.g., visual objects appearing in imagery data) which have been shown to be highly correlated to visual mental imagery [23]. Recently, Garimella et al. [24] proposed to use automatic image tagging to obtain high-level visual information for public health analysis.

2.3. Spatial data

Spatio-referenced data have been utilised in a broad range of applications including ecology and environmental management, transportation, and epidemiology [25]. In public healthcare, accumulated geo-tagged data can be harnessed to determine the health issues, monitor the spread of infectious diseases, and/or analyse the effects of a clinical concept on public health. For instance, geo-tagged data could be used to specify geographic densities of a clinical concept in regions of interest [26, 27], cluster groups of data having similar location characteristics [28], and build recommendation systems to advise locations of interest [29]. The relationship between urban form (e.g., built density, street connectivity, and the amount of green area) and well-being can also be identified from social media [30]. When location information is presented in textual forms, geo-tagging methods can be employed to extract location information from the text [31].

2.4. Social capital data

Online social networks reflect the real-life social networks. People join groups of interest, make friends, show their attitudes and opinions via statuses, comments, likes, and other reactions. The unprecedentedly growth of social media platforms, e.g. Twitter, Facebook, Flicker, etc., has established abundance online connectivities. People interact with others, engage groups of interest, tag friends. As such, social media has made a new form of data, called *digital social capital* (or *social capital* for brevity).

Social capital data has been harnessed to infer social ties from geographic coincidences. This suggests that even a very small number of spatio-temporal co-occurrences can result in a high empirical likelihood of a social tie. Although there still exists no consensus on the definition of social capital, the term generally implies aspects of a social structure and actions made by individuals within the structure [32]. In public health research, social capital is considered as both individual and group attributes [33]. In social science, various instruments have been developed to measure social capital information and empirical studies have shown that different instruments result in different aspects of social capital measures [34].

3. Experimental Design

3.1. Social media data types

3.1.1. Textual data

Most social media-based population health analysis approaches extract the semantical and emotional information from textual content of the media data. In this work, we experiment two common approaches for text analysis including *Bag-of-words* (BOW) and *Linguistic style*.

BOW is a traditional approach for encoding textual data. In this approach, a text (e.g., a document) is represented as a histogram of frequency of words. Words capture the semantical meaning of the textual data and thus could somehow reveal the health status of web users.

Linguistic style is often used in emotion analysis of textual data. This approach relies on associations between language style and health outcomes. For instance, as discovered in [35], depressing stories could also make readers get depressed. Based on these findings, a software package, namely Linguistic Inquiry and Word Count (LIWC), was developed to extract psycho-linguistic features from documents [36]. The LIWC goes through every word of a document, makes comparison of each word with a pre-built dictionary, calculates the percentage of each LIWC category, and finally results in a list of categories occurring in the document.

LIWC has been widely adopted in social media-based health analysis. For instance, Culotta performed a linguistic analysis of activities on Twitter to estimate health indices from County Health Rankings & Roadmaps [37]. Experimental results showed significant correlation (with 6 of the 27 indices) between the language that people used and their health situation. This study also indicated that tweets better captured the health status of a community than demographic. In [38], linguistic features were found to be predictive of the subjective well-being of the U.S. counties.

3.1.2. Visual data

Images from social media data (e.g., images posted/shared by web users) can be considered as visual indicators of a population's health status. To encode this visual data, we propose to extract visual features using convolutional neural networks

(CNNs). CNNs are a specific architecture of feed-forward neural networks designed to exploit the advantages of spatial structures in images. CNNs have demonstrated great performance in understanding imagery content and are being used widely in various research fields such as image classification [39, 40, 41, 42], speech recognition [43, 44], and natural language processing [45, 46].

We employ the CNN architecture proposed in [40]. This network has been trained to classify 1,000 different object classes. Compared with hand-crafted features, e.g., colour histograms, the CNN-based features can be learnt automatically and directly from data. In addition, CNN-based features can represent visual information at different levels of details: from low-level (e.g., pixels) to high-level (e.g., semantical objects).

3.1.3. Social capital data

"Social capital is a compound and complex construct, an umbrella term under which social cohesion, social support, social integration and/or participation are often lumped together" [34]. Social capital has been increasingly appealing policy makers and researchers in public health [34]. Researches in social capital identify two main types of social capital, as summarised in [47]: the *bonding* between individuals in a group and the *bridging* between groups. Each type consists of *structural* and *cognitive* components.

Putnam [48] describes social capital as the presence of "*third places*", i.e., the social surroundings separated from home (referred to as "*first place*") and offices (referred to as "*second place*"). These third places can be cafes, shopping centres, recreational venues, etc., and considered as a broad societal measure of community health. Third places are also vitally considered "*the Heart of a Community*" as they are where humans develop relationships and understanding of society, fill the need for intimacy and affiliation, and earn "*spiritual tonics*" that enrich the their lives [49]. In addition, third places convey structural and cognitive components of both bonding and bridging types of social capital. In other words, third places reflect a community's mental well-being. It is, however, challenging to measure both social capital and mental health precisely.

Inspired by the work in [48], we propose so-called "collective social capital" to

encode the social capital information of counties. Specifically, collective social capital of a county is the measure of the density of the third places in that county across nine categories: Arts & Entertainment, College & University, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, and Travel & Transport. The density of a social venue in a community is contributed by people of that community; the higher the density of a venue is, the more popular the venue is and the more broadly the group/person engaging to that venue could make. Therefore, the presence and density of the third places somehow imply the engagement of people in the community to some interests (events/venues).

3.2. Datasets

We collected user-generated data from two most popular location-centric social networks: Foursquare and Flickr in year 2015. In both the platforms, geographical location of each post, specified by its latitude and longitude coordinates, is available. Foursquare is often used to share user experiences, reviews and recommendations about places on the go, while Flickr is the host for sharing photos and videos. The health indices of the U.S. counties are obtained from the Behavioral Risk Factor Surveillance System. There were 3,105 counties and 342,397,589 tweets collected. The density of social media usage in each county (i.e., the number posts made per person) varied between 0.013 posts/person and 50.939 posts/person, with a mean of 0.64 posts/person and standard deviation of 1.77 posts/person. All counties contributed to the collected data. Fig. 1 shows the distribution of social media usage density of our data.

Foursquare:`https://foursquare.com/about`. . In Foursquare, places are referred to as "venues" and play the central role. At the venues, users can update their check-ins, tastes, likes, tips, recommendations, and purchase history using their mobile phones. A venue is characterised by its name, geographical coordinates, and category. There are 923 categories; a full list of categories can be found at: `https://developer.foursquare.com/docs/resources/categories`. Those categories are organised in a hierarchical structure with 9 categories at the top level. For instance, "Chinese Restaurant" is a sub-category of "Asian Restaurant" which is

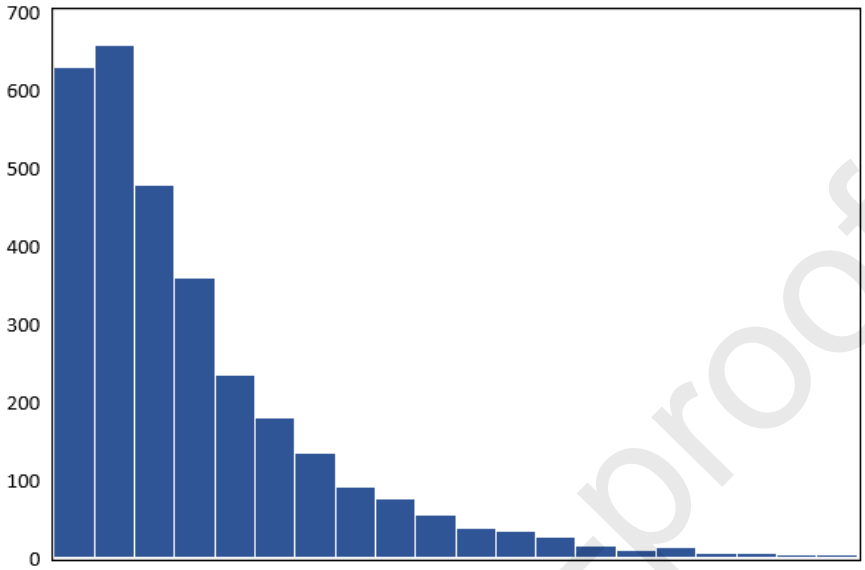


Figure 1: Distribution of density of social media usage. Horizontal axis represents the density of social media usage in counties in the US. Vertical axis represents the number of counties.

under the category "Food" at the top level. For simplicity, in this paper we make use of only the top-level categories, i.e., the top-level category of a venue is considered as the category of that venue. For instance, venues tagged as "Chinese Restaurant" or "Asian Restaurant" will have the category as "Food". Table 1 summarises venues and tips in the top-nine categories of Foursquare collected for this study. Table 2 shows examples of venues and tips of several counties in Foursquare. With more than 50 million monthly active users to date, the network is increasingly becoming a popular large-scale data provider.

Like typical social network platforms, Foursquare also provides its Application Programming Interfaces (APIs) that enable accesses to most of the open data in the network. We collected data from Foursquare venues of the U.S. counties. Each county is encoded by a five-digit Federal Information Processing Standards (FIPS) code, using the cartographic boundary shapefiles provided by the U.S. Census Bureau at `https://www.census.gov/geo/reference/codes/cou.html`. The county geographical bounds, specified by (South-West, North-East), are derived by employing

Table 1: The numbers of venues and tips in Foursquare.

Category	**#Venues**	**#Tips**
Arts & Entertainment	64,550	45,404
College & University	47,586	29,395
Food	95,482	150,554
Nightlife Spot	61,916	53,715
Outdoors & Recreation	102,290	71,430
Professional & Other Places	103,673	45,705
Residence	43,846	11,157
Shop & Service	107,890	101,670
Travel & Transport	82,622	55,001
Total	**709,855**	**564,031**

Table 2: Examples of venues and tips of several counties on Foursquare.

County code	**Top-level category**	**Tips**
36065	Shope & Service	"Sales associates are friendly"
26029	Professional & Other Places	"This is where you go on a rainy day. You may find yourself hoping for rainy days after you visit it once."
48227	Food	"Food is not bad"
29023	Professional & Other Places	"This is where you go to die. The standard of care is appalling!"
13009	College & University	"The staff here are very friendly and welcoming!"

GeoPandas (http://geopandas.org/). We collected 709,855 unique venues in total of 879,295 returned addresses. These numbers indicate a small amount of overlapping data among the counties.

To gather user activities from the collected venues, requests such as who checked-in, tips created by users, photos, likes were sent to the network. In addition, user information and friend lists were crawled. Since the check-in crawling is supported only to self requests, we alternately collected venue check-in information indirectly by venue photos, wherein a property in the response named "checkin" was set if a photo was posted with user check-in.

Flickr:*https://www.flickr.com*. . We collected photos from Flickr by calling its APIs at https://www.flickr.com/services/api/. To crawl photos within the U.S., we filtered the photos by a bounding box (-170.0, 18.0, -60.0, 72.0). Meta-data, including text and geolocation of the images, was also collected. This resulted in a set of 4,213,052 images. Based on the tagged latitude and longitude information associated with each post, all the collected posts were then mapped to the corresponding U.S. counties. Several preprocessing steps were applied to the collected data. Specifically, the text was undergone a basic natural language processing procedure to filter meaningful words. In addition, only images accompanied with non-blank text were kept. Several of our collected images and associated text are shown in Fig. 2.

Behavioral Risk Factor Surveillance System (BRFS): *https://www.cdc.gov/brfss/index.html*. . We used the available survey reports from BRFSS to collect the health indices of counties. BRFS is the largest conducted health survey system, not only in the U.S. but also in the world, with about 400,000 interviews completed each year. The surveys were conducted by the Centers for Disease Control and Prevention (CDC) via questionnaires and on telephones to collect the U.S. residents' health-related data in regarding to their health-related risk behaviours, chronic health conditions, and health outcomes. The questionnaires were categorised into 18 core sections including health status, healthy days, inadequate sleep, chronic health conditions, and 19 optional modules such as healthcare access or social context. For instance, participants would be asked "*Now thinking about your mental health, which includes stress, depression,*

(a) Friendly's Restaurant

(b) A good evening in the woods

(c) Making pizza at the Cooking Block in Redlands!

(d) O what a beautiful day I have found at the end of this week...

(e) It's been a week of colds and yuckiness, so getting out for lunch today felt sooo good

Figure 2: Some images and associated text from Flickr.

and problems with emotions, for how many days during the past 30 days was your mental health not good?" to evaluate healthy days. A full list of questions can be found at `https://www.cdc.gov/brfss/questionnaires/pdf-ques/2014_BRFSS.pdf`.

The BRFSS also reports county health ranking data annually. The ranking information includes i) poor or fair health (i.e., percentage of adults that report fair or poor health), ii) poor physical health days (i.e., the average number of reported physically unhealthy days per month), and iii) poor mental health days (i.e., the average number of reported mentally unhealthy days per month).

3.3. Predictive Model

We evaluated different social media data types in the task of prediction of health indices of the U.S. counties. To predict a health index of a county, linear regression is employed. In particular, let $\mathbf{x}_T(C) \in \mathbb{R}^d$ denote the feature vector extracted from county C and on social media type T (e.g., T="text"). The health index $y_T(C) \in \mathbb{R}$ of county C can be predicted as,

$$y_T(C) = w_0 + \mathbf{w}^\top \mathbf{x}_T(C) + e \tag{1}$$

where $w_0 \in \mathbb{R}$ is the bias, $\mathbf{w} \in \mathbb{R}^d$ is the weight vector, and $e \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian error term.

To train the predictive model, the bias w_0 and weight vector $\mathbf{w}$ are learnt to minimise prediction error on training data.

4. Results and Analysis

In this study, we experiment three social media types including textual, visual and collective social capital data. For textual data, BOW and LIWC are used to extract textual features. Various BOW features are extracted by varying the vocabulary size from 50 to 500 words and the average results are reported. LIWC takes text as input and returns a 78-dimensional psycho-linguistic feature vector. In our experiments, the LIWC features are extracted by using the software in [36].

To extract visual features, each image is undergone the well-known VGG-16 CNN model proposed in [40] and embedded into a 1,000-dimensional feature vector. The CNN model has been trained on the large-scale ImageNet dataset (`www.image-net.org`). This allows the network to capture popular and important features from different applications. The visual features for each county are formed by averaging the visual features extracted from all the images captured in that county.

Collective social capital information is extracted as follows. Let C be a county in which a set of venues $V = \{v_1, v_2, ...\}$ is collected. Let $G = \{g_1, g_2, ..., g_9\}$ be the set of nine categories. The collective social capital information of C is represented by

a 9-dimensional vector $\mathbf{x}(C) = \langle x_1, x_2, ..., x_9 \rangle$ where x_i is the third place density of county C over the category g_i, i.e.,

$$x_i = \frac{\sum_{v \in V} 1(cat(v) = g_i)}{|V|}$$

where $cat(v)$ returns the corresponding category of venue v, and $1()$ is an indicator function, i.e., $1(cat(v) = g_i) = 1$ if $cat(v) = g_i$, and 0 otherwise. The category $cat(v)$ of venue v is determined via posts made in v.

In our experiments, for each county and on each social media data type, we predicted three public health indices including generic, physical, and mental. The values of these indices varied in $8 - 40$ for generic, $2.4 - 5.8$ for physical, and $2.2 - 6.3$ for mental health index. We conducted an across-county evaluation. Specifically, we randomly sampled the dataset with 80% was used for traning and 20% was used for testing. Root mean square error (RMSE) between the actual mental health values provided from CDC and the predicted values was used as the measure of prediction quality.

Table 3: Prediction performance (in RMSE) of various social media data types. The smaller the error is, the better the performance is.

	BOW	**LIWC**	**Visual**	**Collective social capital**
Generic	5.10	4.87	3.97	4.86
Physical	0.77	0.73	0.63	0.70
Mental	0.62	0.57	0.53	0.55

Table 3 represents the prediction performance of the three social media types. Experimental results show that, among the three social media data types, visual data described by CNN features achieved the best performance in predicting all the health indices. Specifically, by using visual data, the generic, physical, and mental health index could be predicted with RMSE of 3.97, 0.63, and 0.53 respectively. The results also show that collective social capital data was the first runner-up with relatively close performance compared to visual data. Textual data encoded by LIWC obtained the third-place while BOW showed the least performance.

To further investigate the results, we analysed regression weights used in regression models, i.e., $\mathbf{w}$ in (1), for all the social media data types. Recall that for visual features, the 1,000 object classes generated by the CNN architecture proposed in [40] were used to describe an image. These features capture semantical objects and somehow represent the visual content from imagery data. We represent the 10 object classes corresponding to the top 10 highest weights in Table 4.

Table 4: The 10 object classes with highest weights. Classes are shown in decreasing order of the weights.

Class ID	**Description**
978	seashore, coast, seacoast, sea-coast
497	church, church building
979	valley, vale
483	castle
460	breakwater, groin, groyne, mole, bulwark, seawall, jetty
746	puck, hockey puck
970	alp
975	lakeside, lakeshore
819	stage
562	fountain

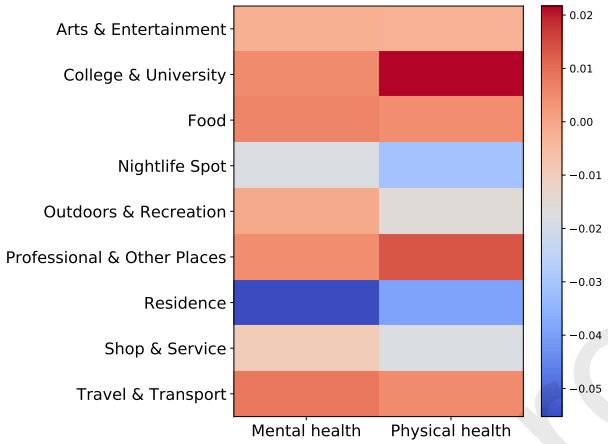


Figure 3: Regression weights of venue categories.

Towards this direction, we also visualised the regression weights of venue categories in collective social capital data and psycho-linguistic features in LIWC. Fig. 3 shows the regression weights of the 9 features of collective social capital data. We found that "Residence", "Outdoor & Recreation", and "Nightlife Spot" were significant indicators of physical health as they were associated with negative coefficients. Mental health was mostly influenced by "Residence" and "Nightlife Spot". Interestingly, "College & University" and "Professional & Other Places" showed strong impact to population well-being, both physical and mental health.

The regression weights of LIWC features are represented in Fig. 4. The weights show that Impersonal Pronouns ("ipron"), Personal Pronouns ("ppron"), and Sexual were most indicative of both mental and physical health. On the other hand, the categories (with positive coefficient): Pronouns, Risk or Death, were weakly associated to the health statuses.

5. Discussion

Although social media data has been used widely for population health analysis, there exist several important issues that need to be addressed.

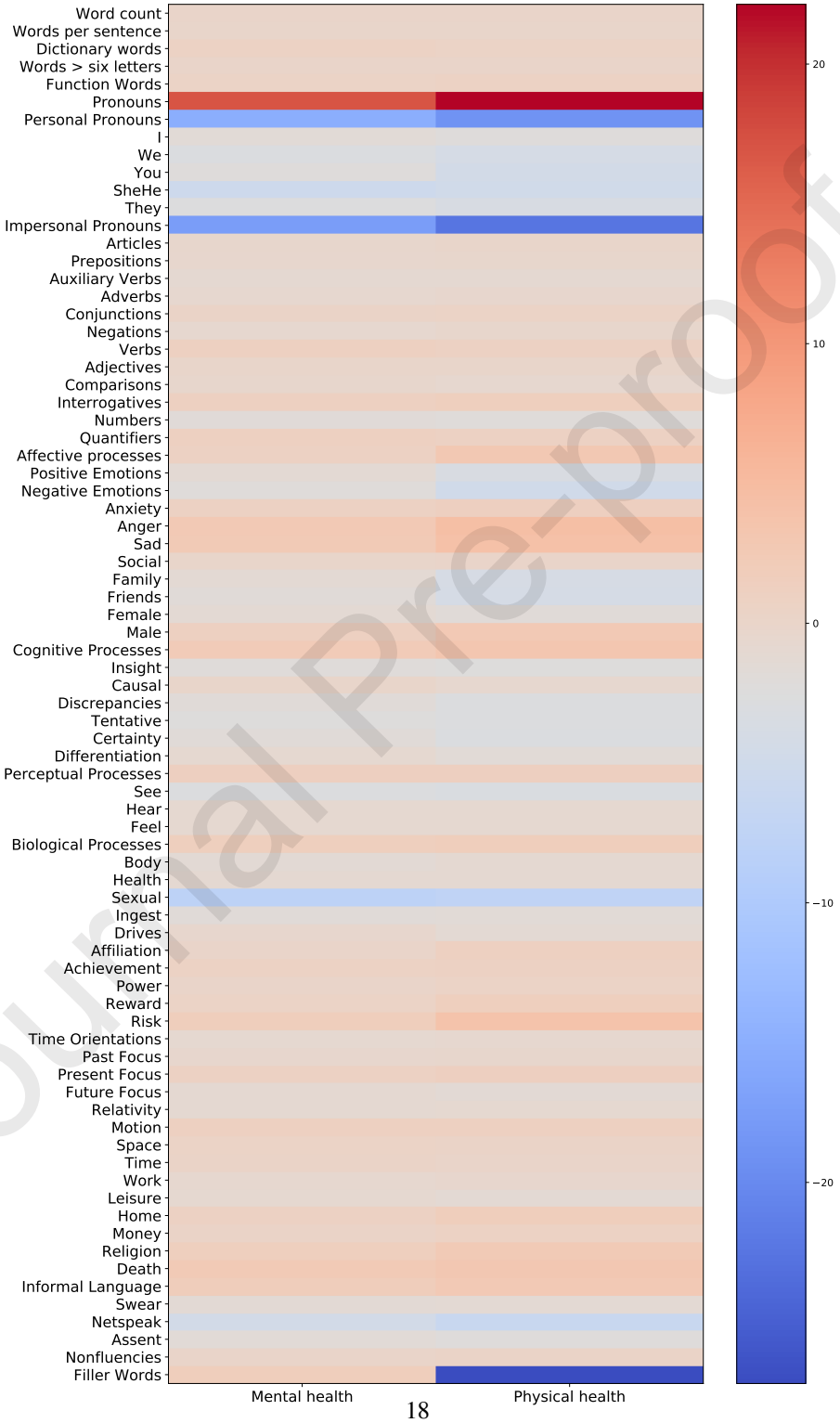


Figure 4: Regression weights of LIWC features.

- There lack of research on the changes of the number of people joining social media in a community and its impact to the overall health of that community. The changes may happen due to physical relocations of the people of the community and/or stopping using social networks though there seems the number of social media users has been keeping increasing. In addition, rural and urban communities may contribute differently to the number of posts and the content of the posts.

- Availability and validity of social media data is an open question for social media-based population health analysis research. Although geocoded information could be used to determine populations and the large-scale factor of the data could compensate the presence of noise (i.e., irrelevant posts), the collected data may be biased by outliers. For instance, posts/pictures in tourist-focused areas may not be made by local people while privacy may be an issue to several social media platforms.

- Temporal changes in social media data are important to be studied. These changes may provide insights in the way people make their posting patterns and relevant temporal health-related issues.

- Social capital can be understood differently [34] and its measurement on online social networks is still debatable [50]. The definition of social capital by Putnam [48] had been proposed before the advent of social networks. Current understandings of social capital in the practice would better fit the current development of social networks and thus should be investigated. More personal aspects such as user mobility and user networking may be important and revealable to personal health status.

- Existing approaches work well in the task of population health prediction but there omits the answer to a fundamental question, that is, what people are implicitly or explicitly concerned in their posts and how these concerns are relevant to their health.

6. Conclusion

This paper conducted an empirical study on the use of social media data for prediction of population health outcomes. In the paper, we studied tree popular types of social media data including text, images, and social capital. We investigated different textual features, proposed the use of visual features learnt from deep neural networks and collective social capital information from location-based social media data. We evaluated the tree types of social media data and relevant features on two popular large-scale datasets: Foursquare and Flickr. Experimental results indicated that visual data and collective social capital data were important indicators of population health outcomes. Exploring more personal aspects of social capital data such as user mobility, user networking, economic, and cultural capital is currently being studied. Combining multi-source/multi-social media data type will also be investigated in our future work.

The work conducted in this paper would be a solution for public health scientists and policy markers to understand the communities' health situation in an automatic yet reliable and scalable manner. Based on the estimated health outcomes, the health status of communities can be understood and appropriate healthcare programs (e.g., health insurance support, mental health education and services) can be determined timely.

References

[1] M. Dredze, How social media will change public health, IEEE Intelligent Systems 27 (4) (2012) 81–84.

[2] R. G. Parrish, Peer reviewed: Measuring population health outcomes, Preventing Chronic Disease 7 (4).

[3] S. B. Thacker, D. F. Stroup, V. Carande-Kulis, J. S. Marks, K. Roy, J. L. Gerberding, Measuring the public's health, Public Health Reports 121 (1) (2006) 14–22.

[4] N. Eagle, M. Macy, R. Claxton, Network diversity and economic development, Science 328 (5981) (2010) 1029–1031.

[5] J. S. House, K. R. Landis, D. Umberson, Social relationships and health, Science 241 (4865) (1988) 540–545.

[6] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), 2013, pp. 128–137.

[7] M. D. Choudhury, S. Sharma, E. Kiciman, Characterizing dietary choices, nutrition, and language in food deserts via social media, in: Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social computing, 2016, pp. 3197–3206.

[8] K. Relia, M. Akbari, D. Duncan, R. Chunara, Socio-spatial self-organizing maps: Using social media to assess relevant geographies for exposure to social processes, in: Proceedings of the ACM Conference on Human-Computer Interaction, 2018, pp. 1–23.

[9] J. Krumm, E. Horvitz, Eyewitness: Identifying local events via space-time signals in twitter feeds, in: Proceedings of the SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2015.

[10] D. Phung, S. K. Gupta, T. Nguyen, S. Venkatesh, Connectivity, online social capital, and mood: A Bayesian nonparametric analysis, IEEE Transactions on Multimedia 15 (6) (2013) 1316–1325.

[11] T. Nguyen, B. Dao, D. Q. Phung, S. Venkatesh, M. Berk, et al., Online Social Capital: Mood, Topical and Psycholinguistic Analysis, in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 2013, pp. 449–456.

[12] M. Salathe, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, et al., Digital epidemiology, PLoS Computational Biology 8 (7) (2012) e1002616.

[13] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (7232) (2009) 1012–1014.

[14] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: Detecting influenza epidemics using Twitter, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1568–1576.

[15] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.

[16] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression in populations, in: Proceedings of the Annual ACM Web Science Conference, 2013, pp. 47–56.

[17] M. J. Paul, M. Dredze, You Are What You Tweet: Analysing Twitter for Public Health, in: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), 2011.

[18] M. J. Paul, M. Dredze, A model for mining public health topics from Twitter, Health 11 (2012) 16–6.

[19] S. Abbar, Y. Mejova, I. Weber, You tweet what you eat: Studying food consumption through twitter, in: Proceedings of the ACM Conference on Human Factors in Computing Systems, 2015, pp. 3197–3206.

[20] M. Akbari, X. Hu, N. Liqiang, T.-S. Chua, From tweets to wellness: Wellness event detection from twitter streams, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2016, pp. 87–93.

[21] L. Manikonda, M. De Choudhury, Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2017, pp. 170–181.

[22] A. G. Reece, C. M. Danforth, Instagram photos reveal predictive markers of depression, EPJ Data Science 6 (1) (2017) 15.

[23] S. M. Roldan, Object recognition in mental representations: Directions for exploring diagnostic features through visual mental imagery, Frontiers in Psychology 8 (833) (2017) 1–15.

[24] V. R. K. Garimella, A. Alfayad, I. Weber, Social media image analysis for public health, in: Proceedings of the ACM Conference on Human Factors in Computing Systems, 2016, pp. 5543–5547.

[25] S. Shekhar, Z. Jiang, R. Y. Ali, E. Eftelioglu, X. Tang, V. Gunturi, X. Zhou, Spatiotemporal data mining: A computational perspective, ISPRS International Journal of Geo-Information 4 (4) (2015) 2306–2338.

[26] U. França, H. Sayama, C. McSwiggen, R. Daneshvar, Y. Bar-Yam, Visualizing the "Heartbeat" of a city with Tweets, Complexity 21 (6) (2016) 280–287.

[27] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, E. Shook, Mapping the global Twitter heartbeat: The geography of Twitter, First Monday 18 (5).

[28] D. Quercia, L. Capra, J. Crowcroft, The social world of Twitter: Topics, geography, and emotions, Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM) 12 (2012) 298–305.

[29] M. Ye, P. Yin, W.-C. Lee, Location recommendation for location-based social networks, in: Proceedings of the SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2010, pp. 458–461.

[30] A. Venerandi, G. Quattrone, L. Capra, City form and well-being: What makes london neighborhoods good places to live?, in: Proceedings of the SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2016.

[31] R. Lan, M. D. Lieberman, H. Samet, The picture of health: Map-based, collaborative spatio-temporal disease tracking, in: Proceedings of the SIGSPATIAL International Workshop on Use of GIS in Public Health, 2012, pp. 27–35.

[32] C. James, Foundations of social theory, Cambridge, MA: Belknap.

[33] I. Kawachi, Commentary: Social capital and health - making the connections one step at a time, International Journal of Epidemiology 35 (4) (2006) 989–993.

[34] I. Kawachi, S. V. Subramanian, D. Kim, Social capital and health, in: Social capital and health, Springer, 2008, pp. 1–26.

[35] J. W. Pennebaker, S. K. Beall, Confronting a traumatic event: Toward an understanding of inhibition and disease, Journal of Abnormal Psychology 95 (3) (1986) 274.

[36] J. W. Pennebaker, R. J. Booth, R. L. Boyd, M. E. Francis, Linguistic Inquiry and Word Count: LIWC 2015 [Computer software], Pennebaker Conglomerates, Inc. (2015).

[37] A. Culotta, Estimating county health statistics with Twitter, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014, pp. 1335–1344.

[38] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. Seligman, L. Ungar, Characterizing geographic variation in well-being using tweets, in: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), 2013, pp. 583–591.

[39] O. Russakovsky, J. Deng, H. Su, et al., ImageNet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[40] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[42] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[43] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu, Convolutional neural networks for speech recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (10) (2014) 1533–1545.

[44] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, The Handbook of Brain Theory and Neural Networks 3361 (10) (1995) 1995.

[45] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the International Conference on Machine Learning, 2008, pp. 160–167.

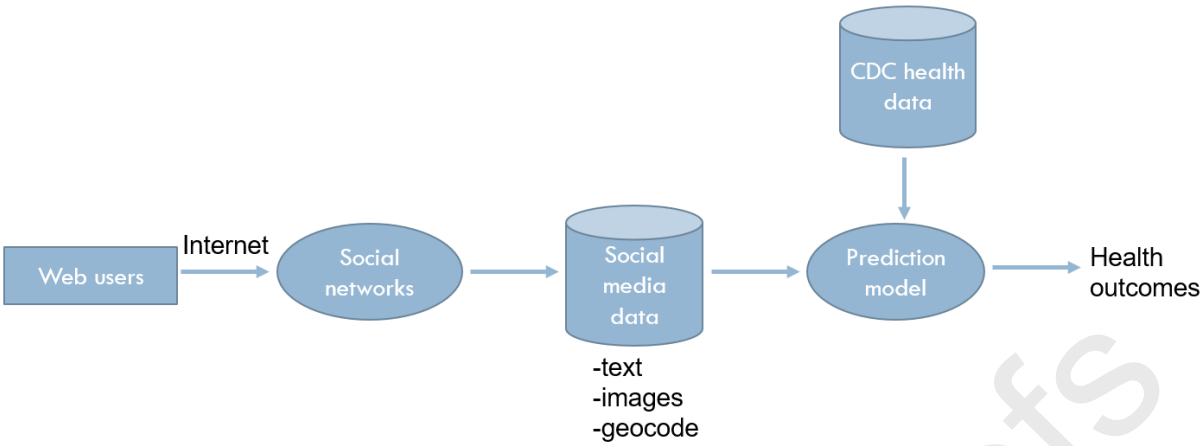
[46] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, Convolutional sequence to sequence learning, ArXiv: 1705.03122.

[47] A. M. Almedom, Social capital and mental health: An interdisciplinary review of primary evidence, Social Science & Medicine 61 (5) (2005) 943–964.

[48] R. D. Putnam, Bowling alone: America's declining social capital, Journal of Democracy 6 (1) (1995) 65–78.

[49] R. Oldenburg, The great good place: Café, coffee shops, community centers, beauty parlors, general stores, bars, hangouts, and how they get you through the day, Paragon House Publishers, 1989.

[50] P. Kazienko, K. Musiał, Social capital in online social networks, in: Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, 2006, pp. 417–424.



We have no conflict of interest.

- Automatic approach for public health analysis using social media data
- Empirical study on various social media data types for population health estimation
- Large-scale experiments and analysis